PERGAMON

Applied Mathematics Letters 16 (2003) 999–1002

Applied Mathematics Letters

www.elsevier.com/locate/aml

Convergence of Online Gradient Methods for Continuous Perceptrons with Linearly Separable Training Patterns

WEI WU* AND ZHIQIONG SHAO
Department of Mathematics
Dalian University of Technology
Dalian 116023, P.R. China
wuweiw@dlut.edu.cn

(Received February 2002; accepted December 2002)

Abstract—In this paper, we prove that the online gradient method for continuous perceptrons converges in finite steps when the training patterns are linearly separable. © 2003 Elsevier Ltd. All rights reserved.

Keywords—Feedforward neural networks, Online gradient method, Convergence, Linearly separable, Continuous perceptrons.

Neural networks have been widely used for solving supervised classification problems. In this paper, we consider the simplest feedforword neural network—the perceptron made up of m input neurons and one output neuron. The objective of training the neural networks is, for a given activation function $g(x) : R^1 \to R^1$, to determine a weight vector $W \in R^m$, such that the training patterns $\{\xi^j\}_{j=1}^J$ are correctly classified according to the output $\zeta = g(W \cdot \xi^j)$ (cf. (2)). Some algorithms training the discrete perceptron where $g(x) = \operatorname{sgn}(x)$, such as the perceptron rule [1] and the delta rule (or Widrow-Hoff rule [2]) based on the LMS (least mean square) algorithm, have proved convergent for linearly separable training patterns. We are concerned in this paper with the continuous perceptron where $g(x)$ is a *sigmoidal* function (a continuous function approximating the sign function $\operatorname{sgn}(x)$). In this case, the online gradient methods are often used for the network training, of which the convergence is our goal in this paper. We expect our analysis here can help to build up similar theories for more important BP neural networks with hidden layers. In this respect, Gori and Maggini [3] prove a convergence result for BP neural networks with linearly separable patterns, under the assumption that the weight vectors keep bounded in the training process. We do not need this restriction in our case.

To train the feedforward neural network (the perceptron), we are supplied with a set of training pattern pairs $\{\xi^j, O^j\}_{j=1}^J \subset R^m \times \{\pm 1\}$, where the ideal output O^j is "1" for a class, and "−1"

Project supported by the National Natural Science Foundation of China, and the Basic Research Program of the National Defence Committee of Science, Technology and Industry of China.
*Author to whom all correspondence should be addressed.

0893-9659/03/$ - see front matter Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX
doi: 10.1016/S0893-9659(03)00135-6

for the other class of patterns. We assume the training patterns are linearly separable, that is, there exists a vector $A \in R^m$ and a constant $C_1 > 0$, such that

$$A \cdot \xi^j = \begin{cases} \geq C_1, & \text{if } O^j = 1, \\ \leq -C_1, & \text{if } O^j = -1. \end{cases} \tag{1}$$

For the purpose of the training iteration, these pairs are arranged stochastically to form a sequence $\{U^k, d^k\}_{k=0}^{\infty} \subset R^m \times \{\pm 1\}$, in which each pair of patterns appears infinite times.

The weight vector to be chosen is $W = (w_1, \ldots, w_m)^\top \in R^m$, where w_j denotes the weight connecting the j^{th} input neuron and the output neuron. For an input vector $U = (u_1, \ldots, u_m)^\top \in R^m$, the output of the network is

$$\zeta = g(h), \qquad h = \sum_{j=1}^{J} u_j w_j = U \cdot W, \tag{2}$$

where $g(x) : R^1 \to I$ $(I = (-1, 1))$ is a given differentiable and bounded activation function. We choose $g(x)$ as *sigmoidal* functions (for example, $g(x) = 2/(1 + \exp(-x)) - 1$). Such a type of function has some important properties, which will be employed in our future proofs, as given below.

PROPERTY 1. $\lim_{x \to \infty} g(x) = 1$, $\lim_{x \to -\infty} g(x) = -1$.

PROPERTY 2. $g(x)$ is an odd function, $g(-x) = -g(x)$.

PROPERTY 3. $\lim_{x \to \pm\infty} g'(x) = 0$.

PROPERTY 4. $\sup_{x \in R} |x g'(x)| = C_0 < \infty$.

PROPERTY 5. $\forall M > 0$, $\exists G_M > 0$, s.t. $g'(x) \geq G_M$ for $-M \leq x \leq M$.

The following properties are direct consequences of the above properties.

PROPERTY 6. $g'(x)$ is an even function, $g'(-x) = g'(x)$ (by Property 2).

PROPERTY 7. $g(x)$ is strictly increasing, so the inverse function $g^{-1}(x)$ exists (by Property 5).

PROPERTY 8. $-1 < g(x) < 1$, $\forall x \in (-\infty, \infty)$ (by Properties 1 and 7).

We train the network to classify the pattern pairs by employing the online gradient method (see, e.g., [4]). So we first select a constant $\varepsilon > 0$ and a random initial weight vector $W^0 \in R^m$. Then at the k^{th} step of the training iteration, we use the input U^k to refine W^k,

$$W^{k+1} = \begin{cases} W^k, & \text{if } \left|d^k - g\left(h^k\right)\right| < \varepsilon, \quad \text{(3a)} \\ W^k + \eta \left(d^k - g\left(h^k\right)\right) g'\left(h^k\right) U^k, & \text{if } \left|d^k - g\left(h^k\right)\right| \geq \varepsilon. \quad \text{(3b)} \end{cases}$$

Next, we perform some simplification and modification of symbols, which proves to be very helpful to our later analysis. First, since what we are really concerned are the actually refined weight vectors W^k in (3b), we can drop out those U^k and W^k that satisfy (3a), and assume every W^k satisfies (3b). Furthermore, if we set $\tilde{\xi}^j = O^j \xi^j$, $\tilde{U}^k = d^k U^k$, then $\{\xi^j, O^j\}_{j=1}^J$ becomes $\{\tilde{\xi}^j, 1\}_{j=1}^J$, and $\{U^k, d^k\}_{k=0}^{\infty}$ becomes $\{\tilde{U}^k, 1\}_{k=0}^{\infty}$. We rewrite ξ^j in place of $\tilde{\xi}^j$, and U^k in place of $\tilde{U}^k$. In these notations, the sequence of input patterns is $\{U^k, 1\}_{k=0}^{\infty}$, and (1) can be revised as

$$A \cdot \xi^j \geq C_1, \qquad j = 1, 2, \ldots, J. \tag{4}$$

And now $\{U^k\}$ and $\{W^k\}$ satisfy

$$1 - g\left(h^k\right) \geq \varepsilon, \tag{5}$$

$$W^{k+1} = W^k + \eta \left(1 - g\left(h^k\right)\right) g'\left(h^k\right) U^k. \tag{6}$$

In fact, by Properties 2 and 6, we see that the weight sequence $\{W^k\}$ remains unchanged under our simplification of symbols. In the sequel, we always assume (4)–(6).

For the training procedure (6), there are two cases to consider.

CASE 1. The training procedure (6) of W^k terminates in a finite number of steps when (3a) is satisfied by a fixed W^k and all the input patterns $\{\xi^j\}_{j=1}^J$.

CASE 2. The training procedure (6) of W^k does not terminate in a finite number of steps and we have an infinite sequence $\{W^k\}_{k=0}^\infty$ satisfies (5) and (6).

We shall proceed by contradiction in the sequel to show that we must have Case 1 to be valid. So until the last theorem we always assume Case 2, or equivalently, assume the existence of the infinite sequence $\{h^k\}_{k=0}^\infty$ satisfying (5) and (6), where $h^k = W^k \cdot U^k$. We remind that the linearly separable condition (4) is now in place of (1).

LEMMA 1. *For the sequence $\{W^k\}_{k=0}^\infty$ defined by (6), there exists a constant $C_1 > 0$ such that*

$$\left\|W^{k+1}\right\|^2 \leq \left\|W^k\right\|^2 + C_1, \qquad k = 1, 2, \ldots. \tag{7}$$

Moreover, there exists a constant $M_1 > 0$, such that for $h^k < -M_1$ we have

$$\left\|W^{k+1}\right\|^2 < \left\|W^k\right\|^2. \tag{8}$$

PROOF. Equation (7) results from he boundedness of $\|U^k\|$ and Properties 4 and 8, by noting

$$\begin{aligned} \left\|W^{k+1}\right\|^2 &= \left\|W^k + \eta\left(1 - g\left(h^k\right)\right) g'\left(h^k\right) U^k\right\|^2 \\ &= \left\|W^k\right\|^2 + 2\eta\left(1 - g\left(h^k\right)\right) g'\left(h^k\right) h^k + \eta^2\left(1 - g\left(h^k\right)\right)^2 g'\left(h^k\right)^2 \left\|U^k\right\|^2. \end{aligned} \tag{9}$$

Observe that $(1 - g(h^k))$ and $g'(h^k)$ are positive and bounded for arbitrary k. Thus, if h^k is small enough, say $h^k < -M_1$ for a suitable constant $M_1 > 0$, there holds

$$2\eta\left(1 - g\left(h^k\right)\right) g'\left(h^k\right) h^k + \eta^2\left(1 - g\left(h^k\right)\right)^2 g'\left(h^k\right)^2 \left\|U^k\right\|^2 < 0.$$

This implies (8) and completes the proof. ∎

In Lemmas 2 and 3 below, we estimate, respectively, the lower and upper bounds of $\{h^k\}_{k=0}^\infty$.

LEMMA 2. *There exists a subsequence $\{h^{k_n}\}_{n=1}^\infty$ of $\{h^k\}_{k=0}^\infty$ and a constant $M_3 \geq M_1$, such that $h^k \geq -M_3$ if $h^k \in \{h^{k_n}\}$, and $h^k < -M_3$ if $h^k \notin \{h^{k_n}\}_{n=1}^\infty$.*

PROOF. We first prove that $h^k \to -\infty$ $(k \to \infty)$ is not possible. We proceed by contradiction. Assume to the contrary that $h^k \to -\infty$ does hold, and then $\forall M_2 \geq M_1$, $\exists K > 0$, such that $h^k < -M_2 \leq -M_1$ for $k > K$. Noting (8), we have $\|W^{k+1}\|^2 < \|W^k\|^2$ when $k > K$; that is, W^k is bounded. So $h^k = W^k \cdot U^k$ is also bounded. But this violates the assumption $h^k \to -\infty$. Thus, $h^k \not\to -\infty$.

The above discussion indicates that $\{h^k\}_{k=0}^\infty$ has a subsequence which is bounded below. Hence, there exists a constant $M_3 > 0$ and a subsequence $\{h^{k_n}\}_{n=1}^\infty$, such that $k_n \to \infty$, and $h^{k_n} \geq -M_3$. Without loss of generality, we may assume $M_3 \geq M_1$ and every h^k which satisfies $h^k \geq -M_3$ is included in this subsequence. This completes the proof. ∎

LEMMA 3. *There exists a constant $M_\varepsilon > 0$ depending on the constant ε in (5), such that $h^k \leq M_\varepsilon$, $\forall\, k = 1, 2, \ldots$.*

PROOF. By the weight updating rule, the weight vector W^k is refined if and only if $1 - g(h^k) \geq \varepsilon$. Therefore, $h^k \leq M_\varepsilon = g^{-1}(1 - \varepsilon) > 0$. ∎

For the weight vector subsequence $\{W^{k_n}\}_{n=1}^\infty$ corresponding to $\{h^{k_n}\}_{n=1}^\infty$, we have the following lemma.

LEMMA 4. *There exists a constant $C_3 > 0$, such that (A is the vector in (1))*

$$A \cdot W^{k_{n+1}} \geq A \cdot W^{k_1} + C_3 n, \qquad \forall n = 1, 2, \ldots. \tag{10}$$

PROOF. Left-multiplying both sides of (6) by A and noting (4) and (5), we derive

$$A \cdot W^{k+1} = A \cdot W^k + \eta \left(1 - g\left(h^k\right)\right) g'\left(h^k\right) A \cdot U^k \geq A \cdot W^k + C_2 g'\left(h^k\right), \tag{11}$$

where $C_2 = \eta \varepsilon C_1$. If $k \notin \{k_n\}_{n=1}^{\infty}$, because $g'(h^k) > 0$, there holds

$$A \cdot W^{k+1} > A \cdot W^k; \tag{12}$$

if $k \in \{k_n\}_{n=1}^{\infty}$, for example $k = k_n$, we conclude from Property 5 and $-M_3 \leq h^{k_n} \leq M_\varepsilon$ that $g'(h^{k_n}) \geq G_{\max\{M_3, M_\varepsilon\}}$. Let $C_3 = C_2 G_{\max\{M_3, M_\varepsilon\}}$. Then (11) implies

$$A \cdot W^{k_n+1} \geq A \cdot W^{k_n} + C_3. \tag{13}$$

It follows from (12) and (13) that

$$A \cdot W^{k_{n+1}} > A \cdot W^{k_{n+1}-1} > \cdots > A \cdot W^{k_n+1} \geq A \cdot W^{k_n} + C_3. \tag{14}$$

This immediately results in (10). ∎

LEMMA 5. *For $\{W^{k_n}\}_{n=1}^{\infty}$, there holds the following estimate:*

$$\left\|W^{k_{n+1}}\right\|^2 \leq \left\|W^{k_1}\right\|^2 + C_1 n, \qquad \forall n = 1, 2, \ldots, \tag{15}$$

where C_1 is the constant in Lemma 1.

PROOF. Very much like the proof of (12)–(14) in Lemma 4, we can derive (15) in terms of (7) and (8) in Lemma 1. The details are omitted. ∎

THEOREM. *For the linearly separable training patterns, the training procedure (5),(6) will converge in finite iteration steps.*

PROOF. Suppose to the contrary that Case 2 is right. Then we have an infinite sequence $\{W^{k_n}\}_{n=1}^{\infty}$ satisfying (10) and (15). By the Schwartz inequality, there holds

$$\|A\| \geq \frac{A \cdot W^{k_{n+1}}}{\|W^{k_{n+1}}\|} \geq \frac{A \cdot W^{k_1} + C_3 n}{\left(\left\|W^{k_1}\right\|^2 + C_1 n\right)^{1/2}} \to \infty, \qquad n \to \infty, \tag{16}$$

leading to a contradiction! So Case 1 must be true, that is, the online gradient method (5),(6) must converge in finite number of iteration steps. ∎

REFERENCES

1. F. Rosenblatt, *Principles of Neurodynamics*, Spartan, New York, (1962).
2. B. Widrow and M.E. Hoff, Adaptive switching circuits, In *Neurocomputing: Foundations of Research*, (Edited by J.A. Anderson and E. Rosenfeld), The MIT Press, Cambridge, MA, (1988).
3. M. Gori and M. Maggini, Optimal convergence of on-line backpropagation, *IEEE Tran. Neural Networks*, 251–154, (1996).
4. W. Wu and Y. Xu, Deterministic convergence of an online gradient method for neural networks, *Journal of Computational and Applied Mathematics* **144** (1/2), 335–347, (2002).